日本国特許庁
JAPAN PATENT OFFICE

31.03.03

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office

出願年月日
Date of Application: ２００２年　４月　４日

出願番号
Application Number: 特願２００２－１０２７４３

[ST.10/C]: [ＪＰ２００２－１０２７４３]

出願人
Applicant(s): 石原産業株式会社

PRIORITY DOCUMENT
SUBMITTED OR TRANSMITTED IN COMPLIANCE WITH RULE 17.1(a) OR (b)

２００３年　５月　９日

特許庁長官
Commissioner,
Japan Patent Office 太田信一郎

出証番号　出証特２００３－３０３３８３５

【書類名】　特許願
【整理番号】　181940
【提出日】　平成14年 4月 4日
【あて先】　特許庁長官殿
【国際特許分類】　G06F 17/00
【発明者】
【住所又は居所】　滋賀県草津市西渋川二丁目３番１号　石原産業株式会社中央研究所内
【氏名】　石川　俊夫
【特許出願人】
【識別番号】　000000354
【住所又は居所】　大阪府大阪市西区江戸堀一丁目３番１５号
【氏名又は名称】　石原産業株式会社
【代理人】
【識別番号】　100062144
【弁理士】
【氏名又は名称】　青山　葆
【選任した代理人】
【識別番号】　100086405
【弁理士】
【氏名又は名称】　河宮　治
【選任した代理人】
【識別番号】　100098280
【弁理士】
【氏名又は名称】　石野　正弘
【手数料の表示】
【予納台帳番号】　013262
【納付金額】　21,000円

【提出物件の目録】

【物件名】　明細書　1

【物件名】　図面　1

【物件名】　要約書　1

【プルーフの要否】　要

【書類名】　明細書

【発明の名称】　データ解析装置および方法

【特許請求の範囲】

【請求項１】　生体の状態と複数の遺伝子発現の量および／または細胞内物質の量との相関モデルを決定するデータ解析装置であって、

生体の状態を目的変数とし、複数の遺伝子発現の量および／または細胞内物質の量を説明変数とするサンプルの集合を入力する入力手段と、

（１）説明変数を選択する選択手段と、（２）部分最小自乗法を実行して交差検証成績を計算する計算手段と、（３）前記（２）の計算手段の結果を評価し、説明変数の採用、不採用を判定する評価判定手段とを有し、（４）前記（１）の選択手段と前記（２）の計算手段と前記（３）の評価判定手段とを実行して部分最小自乗法モデルの交差検証成績を改善し続けて部分最小自乗法モデルを決定する決定手段とからなることを特徴とするデータ解析装置。

【請求項２】　前記の選択手段において、説明変数を逐次取捨選択することを特徴とする請求項１に記載のデータ解析装置。

【請求項３】　前記の選択手段において、遺伝的アルゴリズムを用いて説明変数を選択することを特徴とする請求項１に記載のデータ解析装置。

【請求項４】　前記の計算手段において、１個のサンプルを逐次除外して部分最小自乗法を実行して交差検証成績を計算することを特徴とする請求項１、２または３に記載のデータ解析装置。

【請求項５】　前記の計算手段において、複数のサンプルを逐次除外して部分最小自乗法を実行して交差検証成績を計算することを特徴とする請求項１、２または３に記載のデータ解析装置。

【請求項６】　前記評価判定手段において、前記計算手段の結果から、各計算において除外したサンプルの遺伝子発現から予測される生体の状態を示す目的変数値と、前記除外したサンプルの生体の状態を示す目的変数値との誤差の代表値を求め、当該誤差の代表値が小さくなった場合に、その交差検証成績が改善されたと評価し、説明変数を取捨選択しながら交差検証成績の評価判定を繰り返すことを特徴とする請求項１、２、４または５に記載されたデータ解析装置。

【請求項７】　前記の決定手段において、前記（１）の選択手段と前記(２)の計算手段と前記（３）の評価判定手段とを繰り返し実行して部分最小自乗法モデルの交差検証成績を改善し続けて部分最小自乗法モデルを決定することを特徴とする請求項１、２、３、４、５または６に記載のデータ解析装置。

【請求項８】　前記（１）の選択手段と前記(２)の計算手段とを複数のコンピュータで実行させることを特徴とする請求項１、２、３、４または５に記載のデータ解析装置。

【請求項９】　請求項１で決定された相関モデル及び予測対象のサンプルについて当該モデルにおいて採用された説明変数を入力する入力手段と、入力された該説明変数に基づいて該サンプルの生体の状態を予測判定する予測判定手段からなることを特徴とするデータ解析装置。

【請求項１０】　生体の状態と複数の遺伝子発現の量および／または細胞内物質の量との相関モデルを決定するデータ解析方法であって、

生体の状態を目的変数とし、複数の遺伝子発現の量および／または細胞内物質の量を説明変数とするサンプルの集合を入力する入力ステップと、

(１)説明変数を選択する選択ステップと、(２)部分最小自乗法を実行して交差検証成績を計算する計算ステップと、(３)前記(２)の計算手段の結果を評価し、説明変数の採用、不採用を判定する評価判定ステップとを有し、(４)前記（１）の選択ステップと前記(２)の計算ステップと前記(３)の評価判定ステップとを実行して部分最小自乗法モデルの交差検証成績を改善し続けて部分最小自乗法モデルを決定する決定ステップとからなることを特徴とするデータ解析方法。

【請求項１１】　前記の選択ステップにおいて、説明変数を逐次取捨選択することを特徴とする請求項１０に記載のデータ解析方法。

【請求項１２】　前記の選択ステップにおいて、遺伝的アルゴリズムを用いて説明変数を選択することを特徴とする請求項１０に記載のデータ解析方法。

【請求項１３】　前記の計算ステップにおいて、１個のサンプルを逐次除外して部分最小自乗法を実行して交差検証成績を計算することを特徴とする請求項１０、１１または１２に記載のデータ解析方法。

【請求項１４】　前記の計算ステップにおいて、複数のサンプルを逐次除外

して部分最小自乗法を実行して交差検証成績を計算することを特徴とする請求項１０、１１または１２に記載のデータ解析方法。

【請求項１５】　前記評価判定ステップにおいて、前記計算手段の結果から、各計算において除外したサンプルの遺伝子発現から予測される生体の状態を示す目的変数値と、前記除外したサンプルの生体の状態を示す目的変数値との誤差の代表値を求め、当該誤差の代表値が小さくなった場合に、その交差検証成績が改善されたと評価し、説明変数を取捨選択しながら交差検証成績の評価判定を繰り返すことを特徴とする請求項１０、１１、１３または１４に記載されたデータ解析方法。

【請求項１６】　前記の決定ステップにおいて、前記（１）の選択ステップと前記(２)の計算ステップと前記（３）の評価判定ステップとを繰り返し実行して部分最小自乗法モデルの交差検証成績を改善し続けて部分最小自乗法モデルを決定することを特徴とする請求項１０、１１、１２、１３、１４または１５に記載のデータ解析方法。

【請求項１７】　前記（１）の選択ステップと前記(２)の計算ステップとを複数のコンピュータで実行させることを特徴とする請求項１０、１１、１２、１３または１４に記載のデータ解析方法。

【請求項１８】　請求項１０で決定された相関モデル及び予測対象のサンプルについて当該モデルにおいて採用された説明変数を入力する入力ステップと、入力された該説明変数に基づいて該サンプルの生体の状態を予測判定する予測判定ステップからなることを特徴とするデータ解析方法。

【請求項１９】　生体の状態と複数の遺伝子発現の量および／または細胞内物質の量との相関モデルを決定する、コンピュータにより実行されるデータ解析プログラムであって、

生体の状態を目的変数とし、複数の遺伝子発現の量および／または細胞内物質の量を説明変数とするサンプルの集合を入力する入力ステップと、

（１）説明変数を選択する選択ステップと、（２）部分最小自乗法を実行して交差検証成績を計算する計算ステップと、（３）前記(２)の計算ステップの結果を評価し、説明変数の採用、不採用を判定する評価判定ステップとを有し、（４）前記（

１）の選択ステップと前記(２)の計算ステップと前記(３)の評価判定ステップとを実行して部分最小自乗法モデルの交差検証成績を改善し続けて部分最小自乗法モデルを決定する決定ステップとからなることを特徴とするデータ解析プログラム。

【請求項２０】　前記の選択ステップにおいて、説明変数を逐次取捨選択することを特徴とする請求項１９に記載のデータ解析プログラム。

【請求項２１】　前記の選択ステップにおいて、遺伝的アルゴリズムを用いて説明変数を選択することを特徴とする請求項１９に記載のデータ解析プログラム。

【請求項２２】　前記の計算ステップにおいて、１個のサンプルを逐次除外して部分最小自乗法を実行して交差検証成績を計算することを特徴とする請求項１９、２０または２１に記載のデータ解析プログラム。

【請求項２３】　前記の計算ステップにおいて、複数のサンプルを逐次除外して部分最小自乗法を実行して交差検証成績を計算することを特徴とする請求項１９、２０または２１に記載のデータ解析プログラム。

【請求項２４】　前記評価判定ステップにおいて、前記計算手段の結果から、各計算において除外したサンプルの遺伝子発現から予測される生体の状態を示す目的変数値と、前記除外したサンプルの生体の状態を示す目的変数値との誤差の代表値を求め、当該誤差の代表値が小さくなった場合に、その交差検証成績が改善されたと評価し、説明変数を取捨選択しながら交差検証成績の評価判定を繰り返すことを特徴とする請求項１９、２０、２２または２３に記載されたデータ解析プログラム。

【請求項２５】　前記の決定ステップにおいて、前記（１）の選択ステップと前記(２)の計算ステップと前記（３）の評価判定ステップとを繰り返し実行して部分最小自乗法モデルの交差検証成績を改善し続けて部分最小自乗法モデルを決定することを特徴とする請求項１９、２０、２１、２２、２３または２４に記載のデータ解析プログラム。

【請求項２６】　前記（１）の選択ステップと前記(２)の計算ステップとを複数のコンピュータで実行させることを特徴とする請求項１９、２０、２１、２

２または２３に記載のデータ解析プログラム。

【請求項２７】　請求項１９で決定された相関モデル及び予測対象のサンプルについて当該モデルにおいて採用された説明変数を入力する入力ステップと、入力された該説明変数に基づいて該サンプルの生体の状態を予測判定する予測判定ステップからなることを特徴とするデータ解析プログラム。

【請求項２８】　上記の説明変数の選択において、初期状態では説明変数を全く含まないことを特徴とする請求項２０に記載されたプログラム。

【請求項２９】　上記の説明変数の選択において、初期状態では全説明変数を含むことを特徴とする請求項２０に記載されたプログラム。

【請求項３０】　上記の生体の状態が病気のタイプまたは医療診断の結果であることを特徴とする請求項１９から２９のいずれかに記載されたプログラム。

【請求項３１】　請求項１９から請求項３０のいずれかに記載されたプログラムを記録した、コンピュータにより読み取り可能な記録媒体。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】

本発明は、生体の状態と遺伝子発現の量および／または細胞内物質の量との多変量解析処理に関するものである。

【０００２】

【従来の技術】

２０００年６月のヒトゲノムの解読宣言以降、ゲノムに書かれた遺伝情報がどのように発現して機能しているかのを解明するポストゲノム時代に突入したと言われている。ヒトゲノム計画の進展の中で、ゲノム発現状態を測定する方法論も進展してきた。トランスクリプトーム（ｍＲＮＡ）測定手段としてオリゴヌクレオチドアレイやマイクロチップが知られている。またプロテオーム（蛋白質）測定手段として、以前からある２次元電気泳動に加えて、最近では質量分析の方法が進歩してきた。また抗体チップなどの先進の技術も注目されている。これらの測定技術は、生体の状態パラメータを短時間に一挙に測定できることがそれまでの技術と比較して画期的であるといえる。

【０００３】

遺伝子発現状態を効率的に測定する技術として次のものがあげられる。トランスクリプトーム（ｍＲＮＡの総体）を特定するものとして、基盤に複数種のＤＮＡを担持し、それに相補的なｍＲＮＡを検出するＤＮＡチップが知られている。代表的なＤＮＡチップには、遺伝子チップやＤＮＡマイクロアレイがある。また、プロテオーム（蛋白質の総体）を特定するものには、２次元電気泳動、抗体チップ、質量スペクトルを用いるものがある。またメタボローム(代謝中間体を含めた代謝産物の総体)を測定する手法も質量分析などによって試みられており、進展が見られる。

【０００４】

生体内の細胞の状態は遺伝子産物の発現によってよく記述されるため、従来の診断マーカーでは情報が不足している場面でも、精度のより高い診断が可能になるという期待も出てきている。たとえば、次のような研究があげられる。

【０００５】

Ｐ.Ｏ. Brownらは、ＤＮＡチップによってリンパ腫患者の細胞のトランスクリプトームを測定し、クラスター解析によって悪性と良性のリンパ腫（ＤＬＢＣＬ）を別クラスターに分離した（Nature 403(3), 503-11 (2000)）。しかし、これは因果関係（相関関係）のモデルを得る方法ではなく、どの遺伝子がどの程度重要かを判断できない。

【０００６】

A. Alaiyaらは、２次元電気泳動によって子宮がん患者４０人の細胞のプロテオームを測定し、うち２２人のデータから部分最小自乗法診断モデルを構築し、悪性度を説明した（Int. J. Cancer, 86, 731-36 (2000); Electrophoresis, 21, 1210-17 (2000); 国際公開 WO 00/70340)。その際、全変数モデルにおいて１５５３変数からloadingの大きな１７０変数に限定することによって交差検証成績がよくなり（$Q^2=0.84$）、残り１８患者の深刻度（３段階）を１１／１８の比率で正答した。交差検証法がモデル構築の際の指標になるという考えが表明されている。しかし、この方法では、loadingを得る際にまず全変数モデルが成立しなければならない。また、それ以外の変数選択手法が考案されていない。

【０００７】

J. Khanらは、ＤＮＡチップによって小児がん患者の細胞を測定し、ニューラルネットワークによって悪性度を説明した（Nature Medicine, 7(6), 673-79 (2001)）。小児がん（SRBCT）患者８８人のトランスクリプトーム（６５６７遺伝子）を測定し、うち６３人のデータから主成分分析によって１０次元に圧縮し、次に、人工ニューラルネットワーク診断モデルを構築した。ここで、影響力のある上位遺伝子を交差検証法によって絞り込み、９６遺伝子で最良の成績(100%)を得た。このモデルで残り２５人を予測し、９３～１００％の結果を得た。しかし、この方法でも、影響力を得る際にまず全変数モデルが成立しなければならない。またそれ以外の変数選択手法が考案されていない。１０次元のような少ない変数の場合を扱えるが、変数の数が膨大な場合には適用できない。

【０００８】

【発明が解決しようとする課題】

従来の診断マーカーでは情報が不足している場面でも、遺伝子発現情報を活用することで、より精度（解像度）の高い診断が可能になるという期待も出てきている。遺伝子発現状態の測定結果は、膨大な情報量が得られることが従来にはなかった特徴であり、逆に情報量が多いために、効果的なデータ処理なくしてデータの活用はありえない。したがって、有用な知識を獲得するためには効果的な情報処理が欠かせない。前に説明したように、現状ではクラスター解析を中心とする方法が用いられているが、主成分分析などの方法も採用されている。クラスター解析や主成分分析は、教師付学習方法ではないため、病状の因果関係（相関関係）のモデルを得ることはできない。すなわち、どの遺伝子がどの程度重要かを解析結果から得ることができないのが難点である。一方、部分最小自乗法は次元圧縮とモデルフィットを同時に行なう強力な多変量解析手法であるが、変数の数が膨大になった場合にしばしば有意な結果が得られない事態に直面する。したがって、膨大な遺伝子発現情報などから有用な知識を獲得できるような効果的な情報処理が望まれている。

【０００９】

この発明の目的は、多変量の遺伝子発現情報、細胞内物質情報の効果的な情報

処理を提供することである。

【００１０】

【課題を解決するための手段】

本発明に係るデータ解析装置は、生体の状態と複数の遺伝子発現の量および／または細胞内物質の量との相関モデルを決定するデータ解析装置であって、生体の状態を目的変数とし、複数の遺伝子発現の量および／または細胞内物質の量を説明変数とするサンプルの集合を入力する入力手段と、（１）説明変数を選択する選択手段と、（２）部分最小自乗法を実行して交差検証成績を計算する計算手段と、（３）前記（２）の計算手段の結果を評価し、説明変数の採用、不採用を判定する評価判定手段とを有し、（４）前記（１）の選択手段と前記（２）の計算手段と前記（３）の評価判定手段とを実行して部分最小自乗法モデルの交差検証成績を改善し続けて部分最小自乗法モデルを決定する決定手段とからなる。前記選択手段は、たとえば、説明変数を逐次取捨選択したり、遺伝的アルゴリズムを用いて説明変数を選択する。計算手段は、たとえば、１個のサンプルを逐次除外したり、複数のサンプルを逐次除外して部分最小自乗法を実行して交差検証成績を計算する。評価判定手段は、たとえば、計算手段の結果から、各計算において除外したサンプルの遺伝子発現から予測される生体の状態を示す目的変数値と、前記除外したサンプルの生体の状態を示す目的変数値との誤差の代表値を求め、当該誤差の代表値が小さくなった場合に、その交差検証成績が改善されたと判定し、説明変数を取捨選択しながら交差検証成績の評価判定を繰り返す。決定手段は、たとえば、選択手段と計算手段と評価判定手段とを繰り返し実行して部分最小自乗法モデルの交差検証成績を改善し続けて部分最小自乗法モデルを決定する。また、選択手段と計算手段とを複数のコンピュータで実行させることもできる。こうして、相関モデルを構成するとき、多変量解析において、交差検証成績を基準に最適化させることにより説明変数を取捨選択し、説明変数の数を減らす次元圧縮をして良好なモデルを得る。

【００１１】

本発明に係るデータ解析方法は、生体の状態と複数の遺伝子発現の量および／または細胞内物質の量との相関モデルを決定するデータ解析方法であって、生体

の状態を目的変数とし、複数の遺伝子発現の量および／または細胞内物質の量を説明変数とするサンプルの集合を入力する入力ステップと、（１）説明変数を選択する選択ステップと、（２）部分最小自乗法を実行して交差検証成績を計算する計算ステップと、（３）前記（２）の計算手段の結果を評価し、説明変数の採用、不採用を判定する評価判定ステップとを有し、（４）前記（１）の選択ステップと前記（２）の計算ステップと前記（３）の評価判定ステップとを実行して部分最小自乗法モデルの交差検証成績を改善し続けて部分最小自乗法モデルを決定する決定ステップとからなる。前記選択ステップは、たとえば、説明変数を逐次取捨選択したり、遺伝的アルゴリズムを用いて説明変数を選択する。計算ステップは、たとえば、１個のサンプルを逐次除外したり、複数のサンプルを逐次除外して部分最小自乗法を実行して交差検証成績を計算する。評価判定ステップは、たとえば、計算ステップの結果から、各計算において除外したサンプルの遺伝子発現から予測される生体の状態を示す目的変数値と、前記除外したサンプルの生体の状態を示す目的変数値との誤差の代表値を求め、当該誤差の代表値が小さくなった場合に、その交差検証成績が改善されたと判定し、説明変数を取捨選択しながら交差検証成績の評価判定を繰り返す。決定ステップは、たとえば、選択ステップと計算ステップと評価判定ステップとを繰り返し実行して部分最小自乗法モデルの交差検証成績を改善し続けて部分最小自乗法モデルを決定する。また、選択ステップと計算ステップとを複数のコンピュータで実行させることもできる。

【００１２】

本発明に係るデータ解析プログラムは、生体の状態と複数の遺伝子発現の量および／または細胞内物質の量との相関モデルを決定する、コンピュータにより実行されるデータ解析プログラムであって、生体の状態を目的変数とし、複数の遺伝子発現の量および／または細胞内物質の量を説明変数とするサンプルの集合を入力する入力ステップと、（１）説明変数を選択する選択ステップと、（２）部分最小自乗法を実行して交差検証成績を計算する計算ステップと、（３）前記（２）の計算ステップの結果を評価し、説明変数の採用、不採用を判定する評価判定ステップとを有し、（４）前記（１）の選択ステップと前記（２）の計算ステップと前記（３）の評価判定ステップとを実行して部分最小自乗法モデルの交差検証成績を改

善し続けて部分最小自乗法モデルを決定する決定ステップとからなる。

【００１３】

上記のデータ解析プログラムにおいて、前記選択ステップは、たとえば、説明変数を逐次取捨選択したり、遺伝的アルゴリズムを用いて説明変数を選択する。計算ステップは、たとえば、１個のサンプルを逐次除外したり、複数のサンプルを逐次除外して部分最小自乗法を実行して交差検証成績を計算する。評価判定ステップは、たとえば、計算ステップの結果から、各計算において除外したサンプルの遺伝子発現から予測される生体の状態を示す目的変数値と、前記除外したサンプルの生体の状態を示す目的変数値との誤差の代表値を求め、当該誤差の代表値が小さくなった場合に、その交差検証成績が改善されたと判定し、説明変数を取捨選択しながら交差検証成績の評価判定を繰り返す。決定ステップは、たとえば、選択ステップと計算ステップと評価判定ステップとを繰り返し実行して部分最小自乗法モデルの交差検証成績を改善し続けて部分最小自乗法モデルを決定する。また、選択ステップと計算ステップとを複数のコンピュータで実行させることもできる。さらには、上記の説明変数の選択において、たとえば、初期状態では説明変数を全く含まないか、或いは、初期状態では全説明変数を含むこともできる。

【００１４】

上記のデータ解析プログラムにおいて、上記の生体の状態は、たとえば病気のタイプまたは医療診断の結果である。

【００１５】

また、本発明は、前記で決定された相関モデル及び予測対象のサンプルについて当該モデルにおいて採用された説明変数を入力する入力手段と、入力された該説明変数に基づいて該サンプルの生体の状態を予測判定する予測判定手段からなるデータ解析装置、前記で決定された相関モデル及び予測対象のサンプルについて当該モデルにおいて採用された説明変数を入力する入力ステップと、入力された該説明変数に基づいて該サンプルの生体の状態を予測判定する予測判定ステップからなるデータ解析方法及び前記で決定された相関モデル及び予測対象のサンプルについて当該モデルにおいて採用された説明変数を入力する入力ステップと

、入力された該説明変数に基づいて該サンプルの生体の状態を予測判定する予測判定ステップからなるデータ解析プログラムも包含する。

【００１６】

本発明に係るコンピュータにより読取可能な記録媒体は、上記のいずれかのプログラムを記録する。

【００１７】

【発明の実施の形態】

以下、添付の図面を参照して本発明の実施の形態を説明する。

以下に、選択された生体の状態と遺伝子発現の量および／または細胞内物質の量との相関モデルの決定について説明する。ここで、遺伝子発現の用語は、ｍＲＮＡ発現(トランスクリプトーム)や、ｍＲＮＡによる翻訳の結果として生じる蛋白質(プロテオーム)を含むものとして用いる。また、細胞内物質の量とはここではたとえば、代謝中間体を含めた代謝産物全部であるメタボロームを意味する。たとえば、トランスクリプトーム(ｍＲＮＡ)やプロテオーム(蛋白質)の解析において、各サンプルデータは、生体の状態と遺伝子発現の量などからなる。各サンプルはたとえば１０００個以上の膨大な遺伝子発現の量を含む。生体の状態は、たとえば病気のタイプまたは病気の診断指標であるが、より一般的には生体情報であればよい。「病気の診断指標」には、病気の進行度合いのほか、重篤度、深刻度などの表現で表わされるものも含む。ここで、遺伝子発現の量などの測定データは膨大な情報量からなるので、コンピュータを用いた効率的な多変量解析が必要である。

【００１８】

データ収集において、予めいくつかのサンプルについて生体の状態（たとえば診断指標）を判定し、また、そのサンプルされたものから細胞液を獲得し、その細胞液中の多くの遺伝子産物の発現の量などを測定する。本発明の実施の形態のデータ解析では、こうして得られた遺伝子産物の発現の量などと生体の状態（たとえば診断指標）を入力し、相関モデル（たとえば部分最小自乗法モデル）を得る。ここで、コンピュータによる多変量解析プログラムを用いて、診断指標を目的変数とし、遺伝子発現の量および／または細胞内物質の量を説明変数とする因

果関係型の解析を行なって、各説明変数の重要性や影響度に関する情報を得る。

【００１９】

我々は今回、遺伝子発現による医療診断という分野において、データ解析において、交差検証（cross validation）の成績を最適化するように変数を選択することによって良好な相関モデル（たとえば部分最小自乗法モデル）が得られることを見出した。交差検証法では、手持ちのデータを複数群に分割し、その一部のデータ群（訓練集合）だけを使ってフィットしたモデルを用いて残る別のデータ群（テスト集合）を予測することによって、モデルの予測力を試す。通常の部分最小自乗法（PLS）においては潜在変数の次元選択に交差検証法が用いられているが、ここでは、部分最小自乗法において、潜在変数を１次元に固定し、１以上の入力変数（説明変数）を逐次取捨選択しながら、交差検証成績（たとえば平方和の予測誤差）を最適化した。その結果、全変数を採用した場合には有意な相関モデルを得られなかった場合にも、良好でかつ予測力のある相関モデルが得られることが判明したのである。この交差検証法を用いた変数選択の逐次取捨選択により、安定な相関モデルが得られる。なお、ここでいう「最適化」とは、交差検証成績が、説明変数を取捨選択するための、そのときの解析条件の範囲で、改善がみられなくなるまで改良したことを意味しており、交差検証成績がすべての説明変数の組合せの中で最適なものを見出したという意味ではない。この変数選択手法を用いると、病状を決定する因子を少数に特定し、廉価な診断用材料（ＤＮＡチップ、抗体チップ、ＤＮＡ含有ベクターなど）を設計でき、それ自体独自の価値を持つものである。また、この変数選択手法は、予め設定される各種の変数選択条件と共に運用することが可能である。

【００２０】

上に述べたように、説明変数は、交差検証成績を基準に逐次取捨選択される。説明変数を追加する場合は、その説明変数について、交差検証成績評価が改善されなかったと判定された場合には当該説明変数を除外し、改善されたと判定された場合には当該説明変数を追加する。また、説明変数を除外する場合は、その説明変数について、交差検証成績評価が改善されなかったと判定された場合には当該説明変数を除外せず、改善されたと判定された場合には当該説明変数を除外す

る。ここで、１以上の説明変数を選択した場合に、交差検証成績評価は次のように進める。ｎ個のサンプルからいくつかのサンプルを逐次除外して部分最小自乗法モデルを求め、各モデルにおいて除外したサンプルの遺伝子発現の量から予測される生体の状態を示す目的変数と、除外したサンプルの生体の状態を示す目的変数との各々の誤差の代表値を求める。「代表値」とは、和、平均、最大値、中位値、最頻値などのデータを特徴づける値をいう。そして、当該誤差の代表値が小さくなった場合に、交差検証成績が改善されたと判定し、当該説明変数を追加または削除する。この交差検証成績評価を、説明変数を取捨選択しながら逐次繰り返して、交差検証成績を改善し続ける。改善されなくなれば交差検証成績を最適化したとして説明変数の取捨選択を終了する。その結果、取捨選択により絞り込んだ数の説明変数からなる最適な部分最小自乗法モデルが得られる。

【００２１】

因果関係型の解析手法においてはオーバーフィット（over fitting）を避けるための工夫が必要となる。ここでいうオーバーフィットとは、説明変数が多すぎるためにたまたま予測結果と実績とが一致するものの、本当の相関関係をとらえ損なっているため、モデルフィットに用いたデータ以外に予測能力を持たないことをいう。ここでは、相関モデルとして部分最小自乗法を用いるが、部分最小自乗法は次元圧縮とモデルフィットを同時に行なう強力な多変量解析手法であり、オーバーフィットの問題に比較的強いとされている。しかし遺伝子発現状態解析のように膨大な変数を扱う場合には、有意な結果が得られない事態に直面する。従来技術として説明したAlaiyaやKhanの手法は全変数モデルが有意に成立することを前提としているので、変数の絞込みには一般的には適用できない。これに対し、本発明では、交差検証予測結果を最適にするように変数を絞り込むことにより、オーバーフィットを減らすことができた。また、本発明は、前記Khanの手法とは異なり、主成分分析などの前処理を介さない方法である。従来技術では、説明変数が膨大な場合には、有意なモデルを得ることができないことから、予め、全説明変数を基にたとえば、主成分分析などで次元圧縮する前処理をし、これによって得られた説明変数によって解析する方法が用いられる。しかしながら、この方法では、構成したモデルで予測を行なうためには、モデル構成の基となった

全説明変数が必ず必要となり、たとえば、説明変数が遺伝子発現の量であれば、診断用遺伝子チップに担持する遺伝子としては、モデル構成に用いた遺伝子の全てが必要となるかもしくは別の手法を用いて変数選択することが必要となる。一方、本発明においては、説明変数の選択によって説明変数を絞り込んでいるので、たとえば、説明変数が遺伝子発現の量であれば、診断用遺伝子チップに担持する遺伝子は、選択された説明変数に相当する遺伝子を担持すれば良いことになる。

【００２２】

なお、Todeschiniらは、有機化合物の大気中の分解を予測するため、遺伝的アルゴリズムによって交差検証成績を最適化するように変数選択を行ない、重回帰モデルを得ている（P. Gramatics, V. Consonni & R. Todeschini, Chemosphere 38(5), 1371-78 (1999)）。５３化合物と１７５記述子でモデル構築を行ない（Q^2＝0.79)、７変数が選択され、９８化合物の予測を行なった（Q^2＝0.75)。交差検証成績を最適化するように変数選択を行なっている点では、本実施形態と同様の手法である。しかし、重回帰モデルを採用しているために、結果として選択される変数は少数個にとどまり、複数の遺伝子発現の量および／または細胞内物質の量の解析には適用できない。

【００２３】

本実施形態では、生体の状態と複数の遺伝子発現の量および／または細胞内物質の量との相関モデルを決定するとき、交差検証成績を基準に最適化させるように説明変数を逐次追加・除外することによって、説明変数の数を減らし、良好な相関モデルを得る。このようなアプローチの優位性は、下記の実施例から推測されるように、次のとおりである。

１）病気や生体現象の背後で働いている重要な遺伝子やメカニズムを推定／特定でき、理解が深まる。

２）重要な遺伝子産物や細胞内物質だけに絞った廉価な診断用材料（ＤＮＡチップ、抗体チップなど）の設計が可能になる。

【００２４】

本実施形態では、交差検証成績を最適化するように説明変数を段階的に取捨選

択するが、たとえば具体的には、ステップワイズ(step wise)法に代表される説明変数を選択する選択手段と、リーブ・ワン・アウト(leave-one-out)法に代表される交差検証法に部分最小自乗法を適用して計算する計算手段と、前記計算手段の結果を評価し、説明変数の採用、不採用を判定する評価判定手段とを組合せて用いる。すなわち、m個の説明変数の中から１以上の説明変数を選択し、次いで、部分最小自乗法を実行して交差検証成績を計算し、更に、該計算結果を評価して前記で選択した説明変数の採用、不採用を判定する。前記評価判定では、前記計算手段の結果から、各計算において除外したサンプルの遺伝子発現から予測される生体の状態を示す目的変数値と、前記除外したサンプルの生体の状態を示す目的変数値との誤差の代表値を求め、当該誤差の代表値が小さくなった場合に、その交差検証成績が改善されたと評価し、説明変数の取捨選択を判定する。上記したように、選択手段と計算手段と評価判定手段とを実行して部分最小自乗法モデルの交差検証成績を改善し続けて、その改善がみられなくなるまで改良し、部分最小自乗法モデルを決定する。なお、本実施形態では、サンプルを１個づつ逐次除外している(リーブ・ワン・アウト法)が、その代わりに、複数のサンプルを除外して交差検証成績を評価してもよい(リーブ・ｎ・アウト法)し、また、Khan et al.により用いられた３分割法(three-hold)等の他の方法を用いることもできる。３分割法では、説明変数をランダムにシャッフルして３つのグループに分ける。その中の２つのグループを用いてモデルを構成し、残りの１つのグループでモデルを評価する。また、説明変数の選択方法としてはステップワイズ法、非線形アルゴリズム(たとえば遺伝的アルゴリズムなど)を用いてもよく、変数選択に関して予め何らかの条件が分っていれば、それに応じて探索範囲を限定できる。

【００２５】

次に、データの収集と解析について具体的に説明する。図１は、遺伝子発現解析システムを示す。データ収集のため、予めいくつかのサンプルについて診断指標（たとえば病気のタイプないし進行度合いを含む）を判定し、また、そのサンプルされたものから細胞液を獲得し、ＤＮＡチップを用いてその細胞液中の多くの遺伝子産物の発現の量を測定する。測定には、共焦点型レーザスキャナ（たと

えばAflymetrix社、４２８アレイスキャナ）１０を用いる。吸光度によりｍＲＮＡの量が測定される。このデータ収集は公知の方法である。測定データは、コンピュータ１２に送られ解析される。コンピュータ１２は、ＣＰＵ１４を備えた通常の構成のコンピュータであり、それに接続される記憶装置（たとえばハードディスク装置）１６の記録媒体（たとえばハードディスク）には、測定データ１８や解析ソフト２０が格納される。この解析ソフト２０を用いてデータ１８が解析され、生体の状態と遺伝子発現の量などとの相関モデルが決定される。

【００２６】

図２は、コンピュータ１２により実行される、生体の状態と遺伝子発現の量などとの相関モデルを得るためのデータ解析ソフト２０のフローチャートを示す。まず、相関モデル作成用のデータを入力する（Ｓ１０）。データはたとえばＤＮＡチップを用いて収集したものである。入力データ（サンプル集合）は、それぞれ目的変数（たとえば診断指標）とｍ個（たとえば２０００個）の説明変数（たとえば遺伝子発現の量）からなる。また、場合によっては、上述のデータ(訓練集合)以外に、テスト集合のデータを入力する。ここでテスト集合とは交差検証の評価のためのデータ群を意味するのではなく、モデル決定が終了した後にモデルの予測力をテストするためのデータ群である。

【００２７】

まず、初期設定として、選択された説明変数の数を０とし、交差検証成績CVの最良値CV_0を$-\infty$とする（Ｓ１２）。次に、説明変数の選択を行う。まず、説明変数を指す番号ｉを１とし（Ｓ１４）、第ｉ変数（遺伝子発現の量）を仮に採用して（Ｓ１６）、部分最小自乗法を実行し、交差検証成績CVを計算する（Ｓ１８、図３参照）。ここで、リーブ・ワン・アウト処理を用いる。これは、たとえば５０個のサンプルからなる訓練集合において、１番から５０番の全てを順次１個づつ除いて残りの４９個のサンプルで予測した結果と、その時除いた１個の結果とを比較し、その誤差が大きい場合に、仮に選択した説明変数(第ｉ変数)が適していないと判断する手法である。もし、得られた成績CVが現在の最良値CV_0より最適化されれば（Ｓ２０でＹＥＳ）、第ｉ変数を採用し、かつ、成績CVを新らしい最良値CV_0に更新する（Ｓ２２）。しかし、得られた成績CVが最良値CV_0より

大きくなければ（Ｓ２０でＮＯ）、第ｉ変数を採用しない（Ｓ２４）。そして、ステップＳ１４に戻り、同様の処理を繰り返す。この処理を交差検証成績CVが改善されなくなる（Ｓ２６でＮＯ）まで繰り返す。ここで、相関モデルに採用する説明変数については１つづつ段階的に増加(追加)または減少（除外）して成績CVを評価判定している。すなわち、全体としての合致度合いがよくなるように各説明変数を解析に加えるかどうかを逐次判定しながら、説明変数の取捨選択を行い、これを、全体としての合致度合いがよくならなくなるまで繰り返す。以上の処理で改善があると、ふたたびステップＳ１４の初め(i=１)に戻り、それまでに選択されている説明変数を基に、さらに説明変数の選択を繰り返す。なお、ここではモデルの予測力を判断するために、訓練集合とテスト集合とに予め分割しておいたデータ集合を用いてデータ解析しており、上述の解析は、訓練集合を用いて行なった結果であるので、この結果からテスト集合について予測を行い、実測データとの合致度を評価（Ｓ２８）している。このような評価は必ずしも必要でないが、予測力を判断するには有効である。

【００２８】

図３は、リーブ・ワン・アウト処理を含む交差検証成績CVの計算（図２、Ｓ１８）のフローチャートを示す。ここで、選択された変数について交差検証成績が計算される。まず、PRESSの初期値を０とする（Ｓ１８０）。次に、ｎ個の集合内のサンプルを指す番号ｊを１とし（Ｓ１８２）、第ｊサンプル以外のｎ－１個のサンプルで部分最小自乗法を実行し（Ｓ１８４）、第ｊサンプルの目的変数を予測する（Ｓ１８６）。差の自乗を計算してPRESSに加算する（Ｓ１９０）。次に番号ｊを１増加し（Ｓ１８２）、同様の処理をおこなう。これを番号ｊ＝ｎまで各サンプルについて繰り返す。得られたPRESSは、１個のサンプルを順次除外して計算した予測値と実測値との差の平方和であり、予測誤差を表わす量である。この予測残差自乗和PRESSの符号を変えたものを交差検証成績CVとする（Ｓ１９２）。

【００２９】

本実施形態では、交差検証法を用いて、入力変数（説明変数）を段階的に１つづつ追加・除外しながら、交差検証成績(平方和の予測誤差)を最適化する。ここ

で、説明変数の段階的な追加・除外の内容を理解しやすくするため、以下で、さらに具体的に５つのモデル構築手法について説明する。これらは、説明変数の逐次的な選択の手順が異なる。

【００３０】

図４は、第１のモデル構築手法を示す。データ集合においてどの説明変数も選択されていない状態を初期状態とする（Ｓ１１２）。次に、１番目の説明変数から最後(m番目)の説明変数までの未だ選択されていない説明変数ごとに逐次、その説明変数を選択した場合に交差検証成績が改善するかどうかを、リーブ・ワン・アウト処理を用いた交差検証成績評価ステップ（Ｓ１１８）を繰り返しながら判定し、改善する場合にはその説明変数を追加する（Ｓ１１４～Ｓ１２４）。そのような改善と追加がなくなる（Ｓ１２６でＮＯ）まで、１番目の説明変数から上記逐次判定操作を繰り返す。

【００３１】

さらに詳しく説明すると、まず、初期設定として、選択された説明変数の数ＮＰを０とし、交差検証成績CVの最良値CV_0を－∞とする（Ｓ１１２）。次に、説明変数の選択を行う。まず、変数ｉを１とし（Ｓ１１４）、第ｉ変数を仮に採用する（Ｓ１１６）。ただし、第ｉ変数がすでに採用されていれば（Ｓ１１５でＹＥＳ）、ステップＳ１１４に戻る。次に、部分最小自乗法を実行し、交差検証成績CVを計算する（Ｓ１１８）。ここで、リーブ・ワン・アウト処理を用いる。もし、得られた成績CVが現在の最良値CV_0より最適化されれば（Ｓ１２０でＹＥＳ）、第ｉ変数を採用し、かつ、成績CVを新らしい最良値CV_0に更新する（Ｓ１２２）。しかし、得られた成績CVが最良値CV_0より大きくなければ（Ｓ１２０でＮＯ）、第ｉ変数を採用しない（Ｓ１２４）。そして、ステップＳ１１４に戻り、同様の処理を繰り返す。この処理を交差検証成績CVが改善されなくなる（Ｓ１２６でＮＯ）まで繰り返す。以上の処理で改善があると、ふたたびステップＳ１１４に戻り、新しいループを開始する。ここで、それまでに選択されている変数を基に、さらに変数の選択を繰り返す。こうして、データ集合を用いて選択された変数を用いた相関モデルが得られる。

【００３２】

図５は、第２のモデル構築手法を示す。この手法では、全ての説明変数が選択されている状態を初期状態とする（Ｓ２１２）。次に、１番目の説明変数から最後(m番目)の説明変数までの選択されている説明変数ごとに逐次、その説明変数を除外した場合に交差検証成績が改善するかどうかを、リーブ・ワン・アウト処理を用いた交差検証成績評価ステップ（Ｓ２１８）を繰り返しながら判定し、改善する場合にはその説明変数を除外する（Ｓ２１４～Ｓ２２４）。そのような改善と除外がなくなる（Ｓ２２６でＮＯ）まで、１番目の説明変数から上記逐次判定操作を繰り返す。

【００３３】

さらに詳しく説明すると、まず、初期設定として、選択された説明変数の数ＮＰをｍとし、交差検証成績CVの最良値CV_0を－∞とする（Ｓ２１２）。すなわち、すべての説明変数を選択する。次に、説明変数の選択を行う。まず、変数ｉを１とし（Ｓ２１４）、第ｉ変数を仮に除外する（Ｓ２１６）。ただし、第ｉ変数がすでに除外されていれば（Ｓ２１５でＹＥＳ）、ステップＳ２１４に戻る。部分最小自乗法を実行し、交差検証成績CVを計算する（Ｓ２１８）。ここで、リーブ・ワン・アウト処理を用いる。もし、得られた成績CVが現在の最良値CV_0より最適化されれば（Ｓ２２０でＹＥＳ）、第ｉ変数を除外し、かつ、成績CVを新らしい最良値CV_0に更新する（Ｓ２２２）。しかし、得られた成績CVが最良値CV_0より大きくなければ（Ｓ２２０でＮＯ）、第ｉ変数を除外しない（Ｓ２２４）。そして、ステップＳ２１４に戻り、同様の処理を繰り返す。この処理を交差検証成績CVが改善されなくなる（Ｓ２２６でＮＯ）まで繰り返す。以上の処理で改善があると、ふたたびステップＳ２１４に戻り、新しいループを開始する。ここで、それまでに選択されている変数を基に、さらに変数の選択を繰り返す。こうして、データ集合を用いて選択された変数を用いた相関モデルが得られる。

【００３４】

図６は、第３のモデル構成手法を示す。この手法は、第１と第２の手法の直列的な組合せである。まず、どの説明変数も選択されていない状態を初期状態とする（Ｓ１１２）。次に、１番目の説明変数から最後（ｍ番目）の説明変数までの未だ選択されていない説明変数ごとに逐次、その説明変数を選択した場合に交差

検証成績が改善するかどうかを、リーブ・ワン・アウト処理を用いた交差検証成績評価ステップを繰り返しながら判定し、改善する場合にはその説明変数を追加選択し、そのような改善と追加がなくなるまで１番目の説明変数から上記逐次判定操作を繰り返す（Ｓ１１４～Ｓ１２６）。次に、１番目の説明変数から最後（ｍ番目）の説明変数までの選択されている説明変数ごとに逐次、その説明変数を除外した場合に交差検証成績が改善するかどうかを、リーブ・ワン・アウト処理を用いた交差検証成績評価ステップを繰り返しながら判定し、改善する場合にはその説明変数を除外し、そのような改善と除外がなくなるまで、１番目の説明変数から上記逐次判定操作を繰り返す（Ｓ２１４～Ｓ２２６）。

【００３５】

図７は、第４のモデル構築手法を示す。この手法は、第３の手法の変形である。まず、どの説明変数も選択されていない状態を初期状態とする（Ｓ１１２）。次に、１番目の説明変数から最後(ｍ番目)の説明変数までの未だ選択されていない説明変数ごとに逐次、その説明変数を選択した場合に交差検証成績が改善するかどうかを、リーブ・ワン・アウト処理を用いた交差検証成績評価ステップ（Ｓ１１８）を繰り返しながら判定し、改善する場合にはその説明変数を追加選択する（Ｓ１１４～Ｓ１２４）。そのような改善と追加がなくなる（Ｓ１２６でＮＯ）まで、１番目の説明変数から上記逐次判定操作を繰り返す。次に、１番目の説明変数から最後(ｍ番目)の説明変数までの選択されている説明変数ごとに逐次、その説明変数を除外した場合に交差検証成績が改善するかどうかを、リーブ・ワン・アウト処理を用いた交差検証成績評価ステップ（Ｓ２１８）を繰り返しながら判定し、改善する場合にはその説明変数を除外する（Ｓ２１４～２２４）。そのような改善と除外がなくなる（Ｓ２２６でＮＯ）まで、１番目の説明変数から上記逐次判定操作を繰り返す。上記逐次判定追加改善ステップまたは上記逐次判定除外改善ステップで少なくとも一度改善があれば（Ｓ２２７でＹＥＳ）、ステップＳ１１２に戻り、上記操作(Ｓ１１２～Ｓ２２７)を繰り返す。これを改善がなくなる（Ｓ２２７でＮＯ）までおこなう。

【００３６】

図８は、第５のモデル構築手法を示す。この手法は、第１と第２のスキームの

並列的な組合せである。どの説明変数も選択されていない状態を初期状態とする（Ｓ１１２）。次に、１番目の説明変数から最後（ｍ番目）の説明変数までの説明変数ごとに逐次、その説明変数が選択されていない場合にはその説明変数を選択した場合に交差検証成績が改善するかどうかを、リーブ・ワン・アウト処理を用いた交差検証成績評価ステップ（Ｓ１１８）を繰り返しながら判定し、改善する場合にはその説明変数を追加する（Ｓ１１４～Ｓ１２４）。また、選択する説明変数ごとに、その説明変数がすでに選択されている場合には、その説明変数を除外した場合に交差検証成績が改善するかどうかを、リーブ・ワン・アウト処理を用いた交差検証成績評価ステップ（Ｓ２１８）を繰り返しながら判定し、改善する場合にはその説明変数を除外する（Ｓ２１６～Ｓ２２４）。そのような改善と追加または除外がなくなる（Ｓ１２６でＮＯ）まで、１番目の説明変数から上記逐次判定操作を繰り返す。

【００３７】

次に、第４のモデル構築手法(図７)を適用した場合を、表１のデータ集合を例として説明する。このデータ集合に対して、部分最小自乗法による解析を用いて相関モデルを求める。表１のデータでは、サンプルの数ｎは１０であり、また、説明を容易にするため、説明変数の数ｍは１９と少なくしている。表１において、p1は目的変数を表わし、p2～p20は説明変数を表わす。(ただし表１では、表示の便宜のため、p16以降のデータを省略している。)第４手法(図７)のステップＳ１１４、Ｓ２１４とは異なり、説明変数を表わすｉはp20からp2まで逆に逐次処理することとした。CV評価値としてここでは予測残差自乗和(PRESS)を採用した。PRESSが小さいほど、CV評価値はよい。初期状態では、採用された説明変数の数ＮＰは０であり、PRESS=∞ (CV_0=-∞)である。

【００３８】

【表１】

表１　１０個のサンプルのデータ

#	p1	p2	p3	p4	p5	p6	p7	p8	p9	p10	p11	p12	p13	p14	p15
1	0.713	0.105	0.782	0.425	0.164	0.023	0.696	0.543	0.333	0.691	0.336	0.668	0.017	0.061	0.5
2	0.133	0.009	0.071	0.002	0.793	0.872	0.092	0.391	0.63	0.241	0.517	0.369	0.166	0.841	0.1
3	0.545	0.193	0.765	0.334	0.109	0.538	0.578	0.652	0.38	0.501	0.729	0.91	0.865	0.389	0.8
4	0.752	0.915	0.472	0.999	0.798	0.363	0.622	0.487	0.353	0.967	0.778	0.484	0.517	0.982	0.0
5	0.9	0.407	0.534	0.816	0.806	0.42	0.572	0.957	0.12	0.696	0.833	0.051	0.377	0.849	0.4
6	0.455	0.587	0.721	0.53	0.252	0.434	0.882	0.486	0.741	0,243	0.893	0.947	0.462	0.952	0.2
7	0.427	0.652	0.515	0.426	0.764	0.592	0.595	0.595	0.551	0.606	0.416	0.163	0.316	0.718	0.6
8	0.042	0.902	0.274	0,899	0.402	0.469	0.668	0.945	0.746	0.912	0.97	0.515	0.368	0.514	0.4
9	0.935	0.276	0.936	0.101	0.54	0.356	0.899	0.71	0.924	0.792	0.486	0.329	0.501	0.076	0.5
10	0.54	0.021	0.505	0.224	0.724	0.431	0.071	0.968	0.482	0.322	0.773	0.543	0.353	0.107	0.9

【００３９】

【表２】

表２　表１のデータについての１０の段階での変数選択結果

0			∞	－
1	追加	p20	0.111	p20
2	追加	p18	0.090	p18 & p20
3	追加	p16	0.073	p16 & p18 & p20
4	追加	p10	0.073	p10 & p16 & p18 & p20
5	追加	p6	0.062	p6 & p10 & p16 & p18 & p20
6	追加	p3	0.060	p3 & p6 & p10 & p16 & p18 & p20
7	追加	p12	0.055	p3 & p6 & p10 & p12 & p16 & p18 & p20
8	除外	p20	0.053	p3 & p6 & p10 & p12 & p16 &
9	除外	p10	0.050	p3 & p6 & p12 & p16 & p18
10	追加	p13	0.048	p3 & p6 & p12 & p13 & p16 & p15

【００４０】

先に述べたように、変数はp20からp2まで逆の順で処理する。表２は、表１のサンプルについて、左端の数字は、変数の取捨選択で改善がみられた１０の段階を示す。なお、０は初期状態を意味する。次の列の「追加」と「除外」は、追加のループと除外のループの処理であることを意味する。次の列の変数は、追加または除外された変数を示す。次の列は、交差検証成績(PRESSをサンプル数で割ったもの)を示す。右端の列は、その段階で選択されている変数を示す。

【００４１】

初期状態では、変数は全くない状態であり、PRESSは∞である。表２に示すように、最初、p20を説明変数として採用すると、PRESS=0.111となり、初期値に比べて改善されるので、説明変数p20の追加を実施する。次に、変数p19を加えてp19とp20の２つを説明変数とすると、PRESS=0.129となり改善をもたらさないので、p19は追加しない。次に、説明変数p18を加えるとPRESS=0.090となり、改善するので、p18を追加し、p18とp20を説明変数とする。以下同様に表２に示すように続く。(ここで、p10を追加採用するのは、小数点以下４桁目で改善されているためである。)説明変数p20～p2の１回目のループを終了した時点で、説明変数がp3、p6、p10、p16、p18およびp20となり、PRESS=0.60となる。２回目のループでは、説明変数p12が追加され、PRESS=0.55となる。３回目のループでは追加による改善がなく、ひとまずＳ１１４～Ｓ１２６の追加処理を終了し、Ｓ２１４に移る。この時点での部分最小自乗法のフィットならびにリーブ・ワン・アウト予測状況は表３のとおりである。

【００４２】

表３は、１０のサンプルについて、表２の７で示す段階まで処理が進んだ時点での部分最小自乗法のフィットならびにリーブ・ワン・アウト予測状況を示す。ここで、モデル予測とリーブ・ワン・アウト予測のそれぞれにおいて、計算値と実測値との誤差を示す。さらに、その下側に、誤差の自乗平均、相関係数Rの自乗および予測相関係数Qの自乗を示す。

【００４３】

【表３】

表３　表２の段階７での処理結果

		モデル予測値		ﾘｰﾌﾞﾜﾝｱｳﾄ予測	
#	実測値	計算値	誤差	計算値	誤差
1	0.713	0.757	-0.044	0.693	0.020
2	0.133	-0.056	0.189	-0.051	0.184
3	0.545	0.497	0.048	0.480	0.065
4	0.752	0.646	0.106	0.495	0.257
5	0.900	0.687	0.214	0.557	0.343
6	0.455	0.489	-0.034	0.512	-0.057
7	0.427	0.624	-0.198	0.672	-0.245
8	0.042	0.349	-0.307	0.517	-0.475
9	0.935	0.865	0.070	0.782	0.153
10	0.154	0.197	-0.044	0.285	-0.132
	0.093		0.024		0.055

R^2=0.744　　Q^2=0.407

【００４４】

次に、Ｓ２１４から始まる除外処理の１回目のループにおいて、説明変数p10とp20を除外することが改善をもたらした。２回目のループでは改善がなく、Ｓ２１４～Ｓ２２６を終了するが、Ｓ２２７の判断により再度Ｓ１１２に戻る。次に、追加処理の１回目のループにおいて、p13の追加だけが改善をもたらしたが、続く除外処理の１回目のループでは、改善がなかった。もう一度Ｓ１１２に戻り、ステップＳ１１４～Ｓ１２６およびステップＳ２１４～Ｓ２２６では改善がなくなったのを確認して、処理を終了した。こうして選択された説明変数は、p3、p6、p12、p13、p16およびp18の５個であり、PRESS=0.048となった。詳細は表４のとおりである。

【００４５】

表４は、表２の段階１０まで処理が進んだ時点での部分最小自乗法のフィットならびにリーブ・ワン・アウト予測状況を示す。

【００４６】

【表４】

表４　表２の段階１０での処理結果

		モデル予測		リーブワンアウト予測	
#	実測値	計算値	誤差	計算値	誤差
1	0.713	0.771	−0.058	0.663	0.050
2	0.133	−0.013	0.146	0.041	0.092
3	0.545	0.610	−0.065	0.595	−0.050
4	0.752	0.524	0.228	0.380	0.372
5	0.900	0.696	0.205	0.543	0.357
6	0.455	0.591	−0.137	0.623	−0.168
7	0.427	0.638	−0.211	0.696	−0.269
8	0.042	0.189	−0.147	0.268	−0.226
9	0.935	0.841	0.094	0.756	0.179
10	0.154	0.209	−0.055	0.294	−0.140
	0.093		0.022		0.048
		R^2=0.765		Q^2=0.482	

【００４７】

なお、説明変数の数が多い時に強いとされる部分最小自乗法であるが、p20～p2の全てを説明変数として採用した場合には、表５に示すように、PRESS=0.124となった。すなわち、リーブ・ワン・アウト処理は、平均値からの誤差(0.093)よりも悪い成績をもたらす。

【００４８】

【表５】

表５　全ての説明変数を採用した場合の処理結果

		モデル予測		リーブ・ワンアウト予測	
#	実測値	計算値	誤差	計算値	誤差
1	0.713	0.712	0.001	0.527	0.186
2	0.133	-0.073	0.206	0.222	-0.090
3	0.545	0.561	-0.016	0.538	0.007
4	0.752	0.656	0.096	0.351	0.402
5	0.900	0.691	0.209	0.432	0.469
6	0.455	0.519	-0.064	0.562	-0.107
7	0.427	0.583	-0.156	0.629	-0.203
8	0.042	0.430	-0.388	0.724	-0.682
9	0.935	0.794	0.140	0.480	0.454
10	0.154	0.182	-0.029	0.457	-0.303
		0.093	0.029		0.124
		R^2=0.684		Q^2=-0.330	

【００４９】

次に、実施例について説明する。P.O. Brownらのホームページ（http://llmpp.nih.gov/lymphoma/）より入手した２８名のＤＬＢＣＬ（リンパ腫）患者のデータを、２０名のデータからなる訓練集合と８名のデータからなるテスト集合に分けた。目的変数に生存月数を採用し、説明変数には１８４３２スポットのうち、２８データにおいてch1、ch2ともに正の数となる１２８３２スポットのlog(ch1/ch2)値を採用した。

【００５０】

訓練集合において部分最小自乗法（ＰＬＳ）のモデル決定を試みた。１２８３２変数全てを用いて部分最小自乗法の解析をしたところ、リーブ・ワン・アウト予測は有意(Q^2＞０.５)にはならなかった。次にリーブ・ワン・アウト予測誤

差が最小になるように説明変数を段階的に１つづつ増減した。モデル構成手法としては前述の第３のモデル構成手法において説明変数の追加及び除外の順番並びにリーブ・ワン・アウト処理におけるサンプルの除外の順番が異なるほかは同様な方法を用いた。すなわち、どの説明変数も選択されていない状態を初期状態とする（Ｓ１１２）。次に、最後（m番目）の説明変数から最初（１番目）の説明変数までの未だ選択されていない説明変数ごとに逐次、その説明変数を選択した場合に交差検証成績が改善するかどうかを、リーブ・ワン・アウト処理（ここでは、最後（ｎ番目）のサンプルから最初（1番目）のサンプルを逐次除外した）を用いた交差検証成績評価ステップを繰り返しながら判定し、改善する場合にはその説明変数を追加選択し、そのような改善と追加がなくなるまでm番目の説明変数から上記逐次判定操作を繰り返す（Ｓ１１４～Ｓ１２６）。次に、最後（m番目）の説明変数から最初（１番目）の説明変数までの選択されている説明変数ごとに逐次、その説明変数を除外した場合に交差検証成績が改善するかどうかを、リーブ・ワン・アウト処理｛ここでも最後（ｎ番目）のサンプルから逐次除外した｝を用いた交差検証成績評価ステップを繰り返しながら判定し、改善する場合にはその説明変数を除外し、そのような改善と除外がなくなるまで、最後（m番目）の説明変数から上記逐次判定操作を繰り返す（Ｓ２１４～Ｓ２２６）。その結果、有意なモデル（R^2＝0.988、Q^2＝0.895、Np＝342）を得た。図９は、このデータについての最小自乗法成績を示す。図９において、ひし形は訓練集合のデータ（20人）を示し、３角は、それらについての交差検証成績のデータを示す。また、４角はテスト集合のデータ（８人）を示す。得られた部分最小自乗法モデルは、テスト集合のうち、４/８をきわめて良好に、また１/８を良好に予測するものであった。

【００５１】

なお、上述の多変量解析によるデータ解析では、扱ったサンプルはＤＮＡチップを用いて得たデータであった。しかし、このデータ解析は、ＤＮＡチップを用いて得たデータに限定されるものではなく、蛋白質発現量、細胞内物質の量などのデータに対しても有用であろうことは容易に推測されることである。

【００５２】

【発明の効果】

生体の状態と複数の遺伝子発現の量および／または細胞内物質の量との相関モデルを決定するとき、説明変数の選択と交差検証法とを用いて変数を絞り込むことができる。これにより、良好でかつ予測力のある多変量解析モデル（相関モデル）が得られる。特に遺伝子発現の量のように、説明変数の数がたとえば１０００以上と膨大な場合に有用である。変数の数を少なくすることにより、病気や生体現象の背後で働いている重要な遺伝子やメカニズムを推定／特定でき、理解が深まる。また、重要な遺伝子産物や細胞内物質だけに絞った廉価な診断用材料（ＤＮＡチップ、ＤＮＡ含有ベクター、抗体チップなど）を設計し、提供できる。

【図面の簡単な説明】

【図１】　遺伝子発現解析システムのブロック図

【図２】　解析ソフトのフローチャート

【図３】　交差検証成績CVの計算のフローチャート

【図４】　変数選択の第１モデル構築手法のフローチャート

【図５】　変数選択の第２モデル構築手法のフローチャート

【図６】　変数選択の第３モデル構築手法のフローチャート

【図７】　変数選択の第４モデル構築手法のフローチャート

【図８】　変数選択の第５モデル構築手法のフローチャート

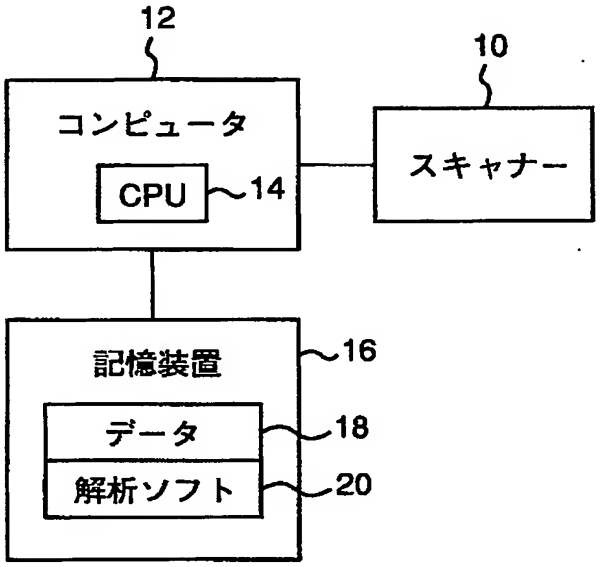
【図９】　最小自乗法モデルの成績を示すグラフ

【符号の説明】

１０　スキャナ、　１２　コンピュータ、　１４　ＣＰＵ、　１６　記憶装置、　１８　記録媒体、　２０　解析ソフト。

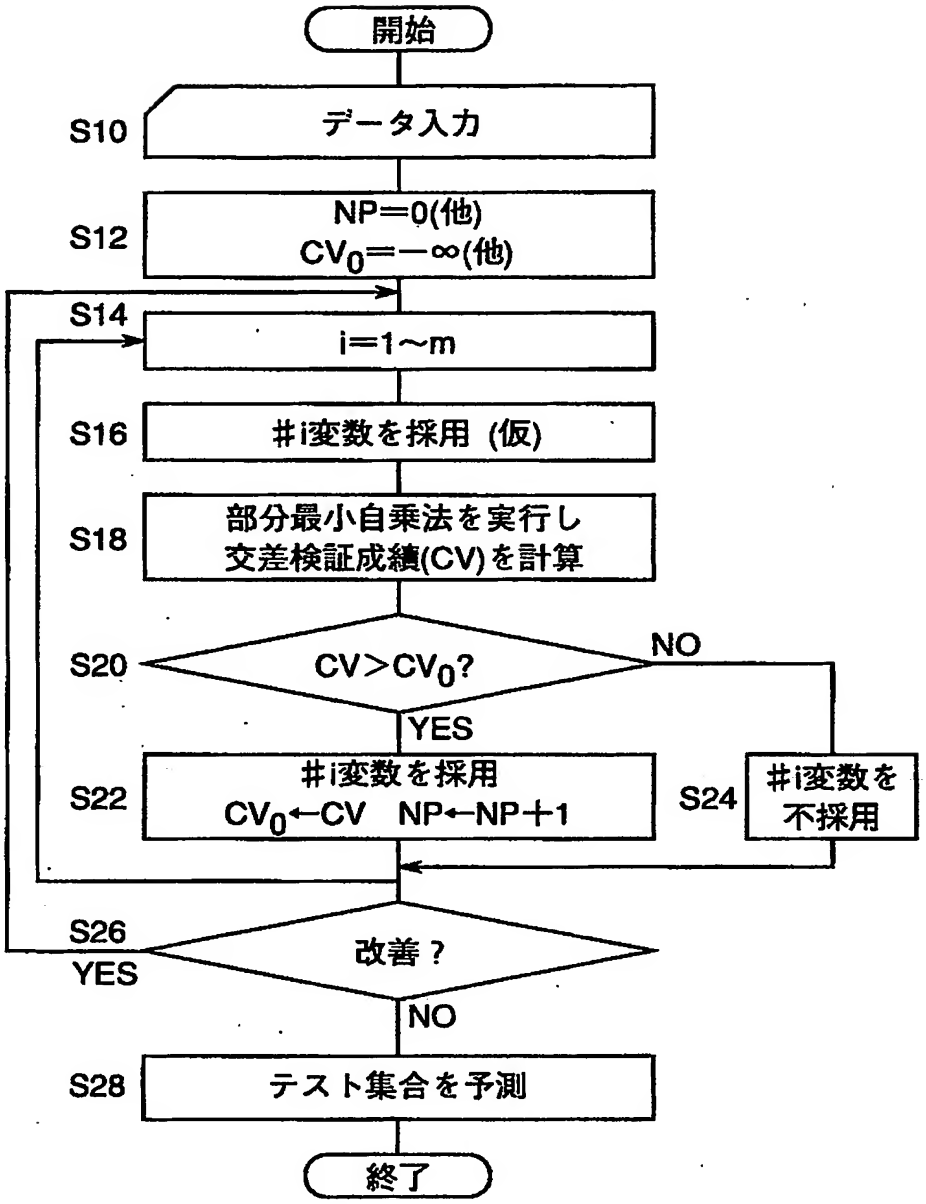
【書類名】　図面

【図１】

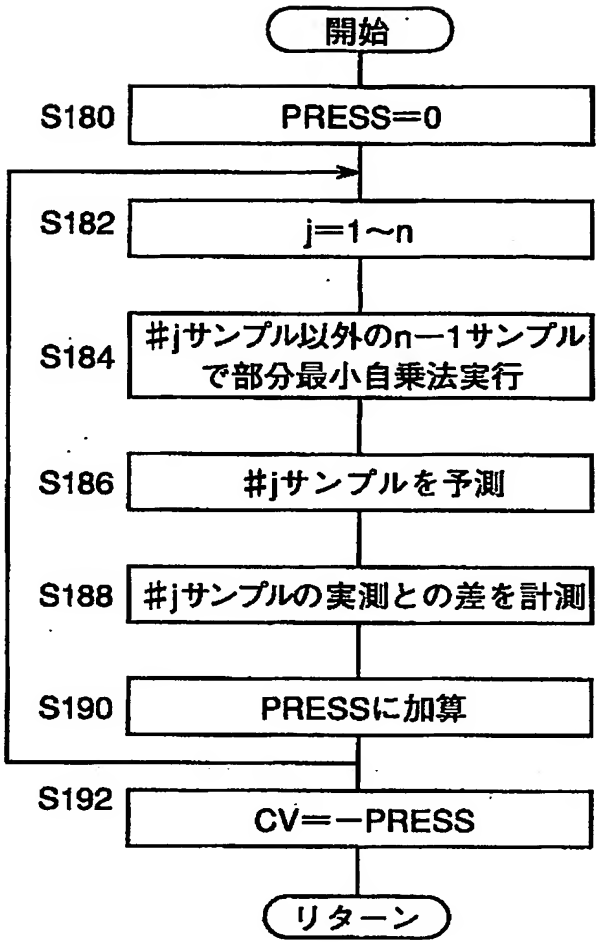


出証特２００３－３０３３８３５

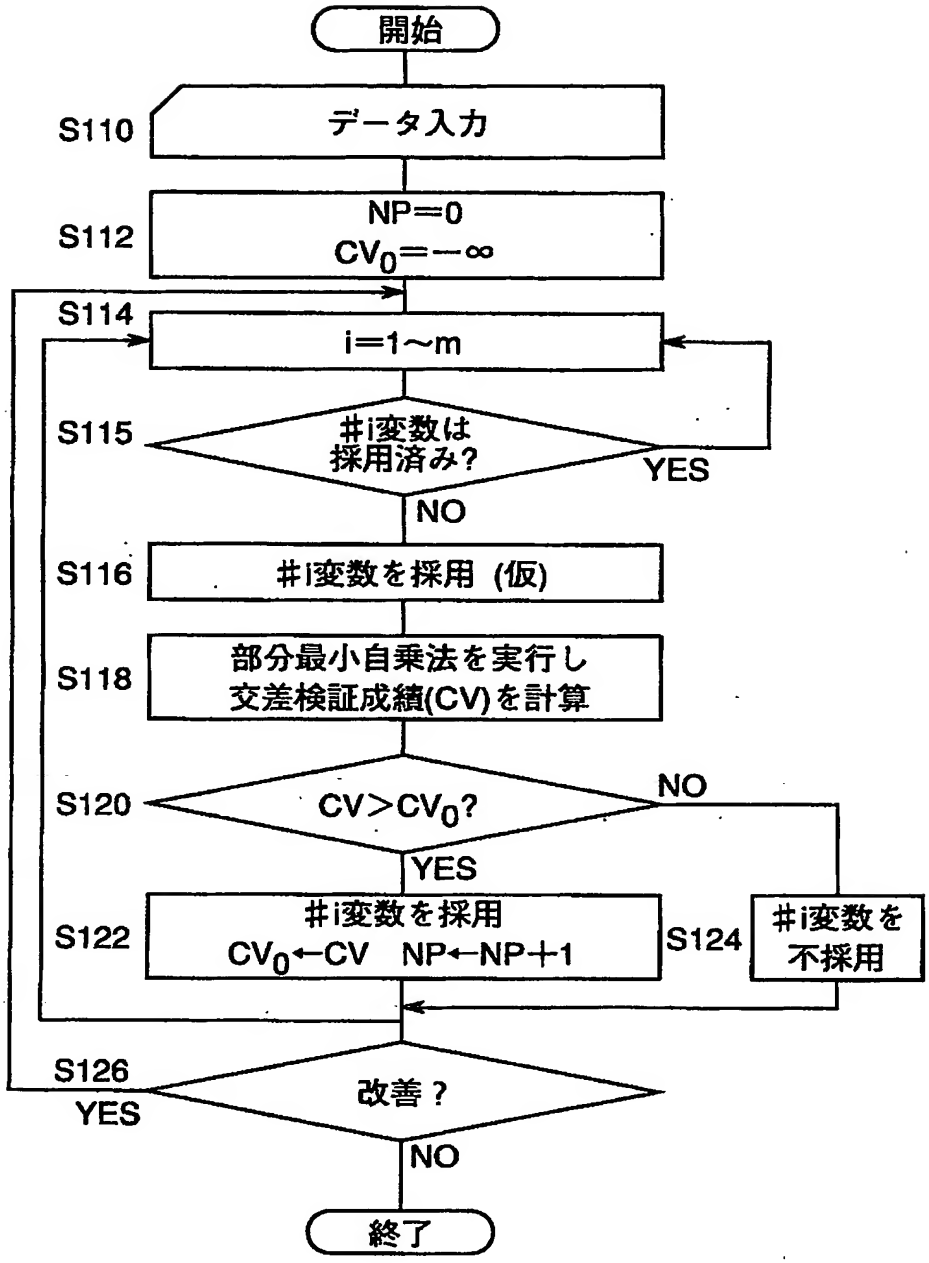
【図２】


開始
S10
データ入力
S12
NP=0(他)
CV0=−∞(他)
S14
i=1～m
S16
#i変数を採用 (仮)
S18
部分最小自乗法を実行し
交差検証成績(CV)を計算
S20
CV>CV0?
NO
YES
S22
#i変数を採用
CV0←CV NP←NP+1
S24
#i変数を
不採用
S26
改善？
YES
NO
S28
テスト集合を予測
終了

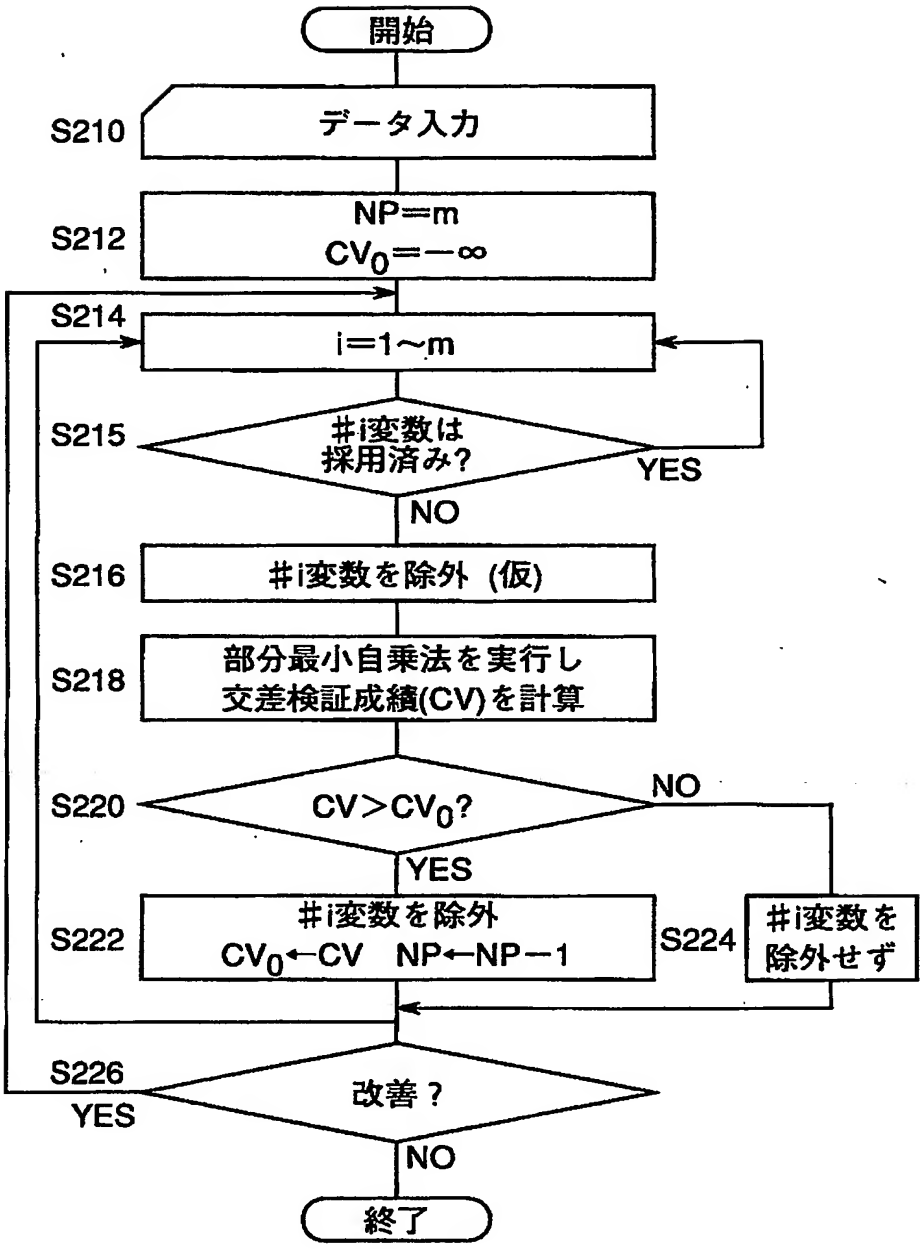
【図３】


開始
S180
PRESS=0
S182
j=1～n
S184
#jサンプル以外のn－1サンプルで部分最小自乗法実行
S186
#jサンプルを予測
S188
#jサンプルの実測との差を計測
S190
PRESSに加算
S192
CV=－PRESS
リターン

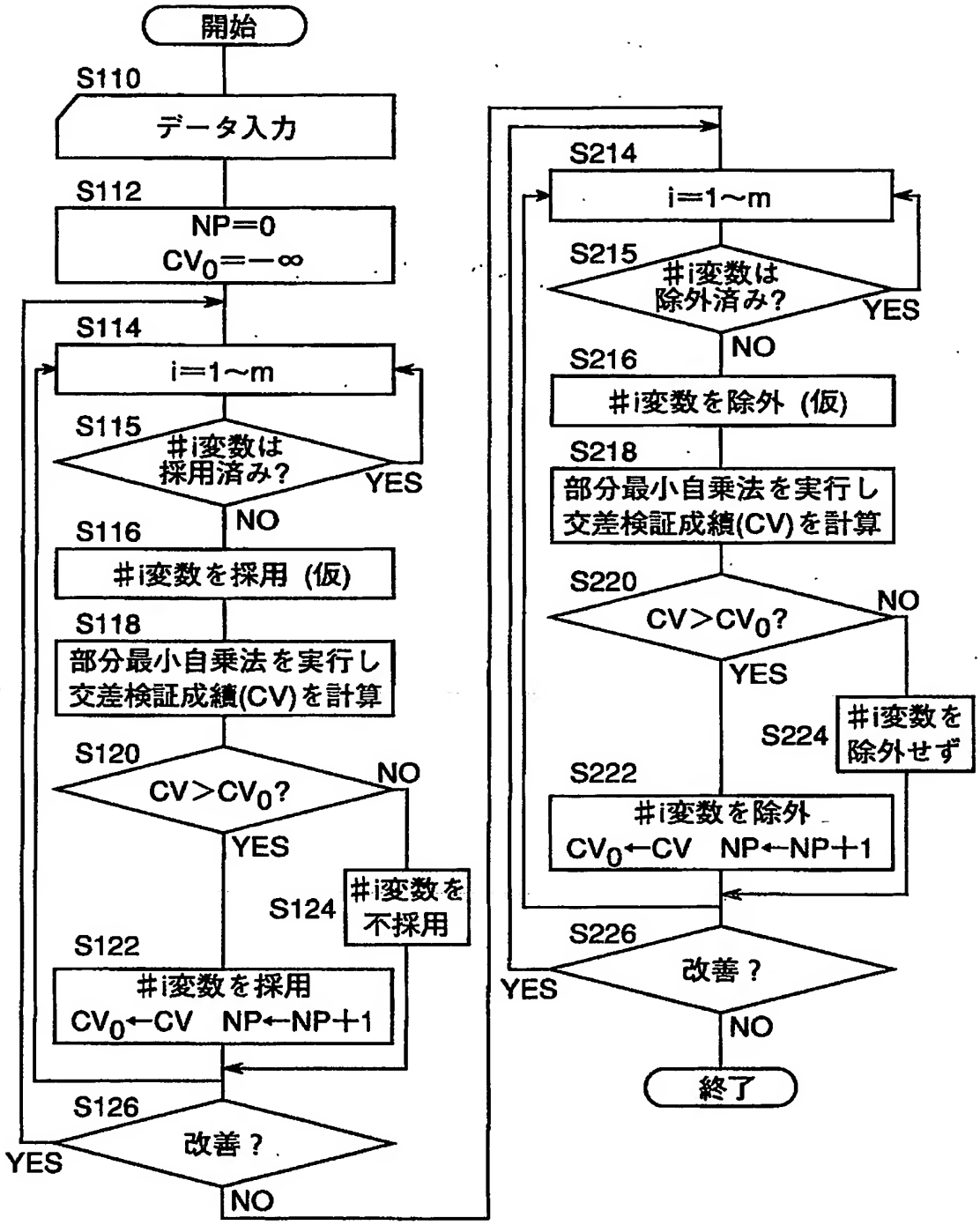
【図４】


開始
S110
データ入力
S112
NP=0
CV0=−∞
S114
i=1～m
S115
#i変数は
採用済み?
YES
NO
S116
#i変数を採用 (仮)
S118
部分最小自乗法を実行し
交差検証成績(CV)を計算
S120
CV>CV0?
NO
YES
S122
#i変数を採用
CV0←CV NP←NP+1
S124
#i変数を
不採用
S126
改善？
YES
NO
終了

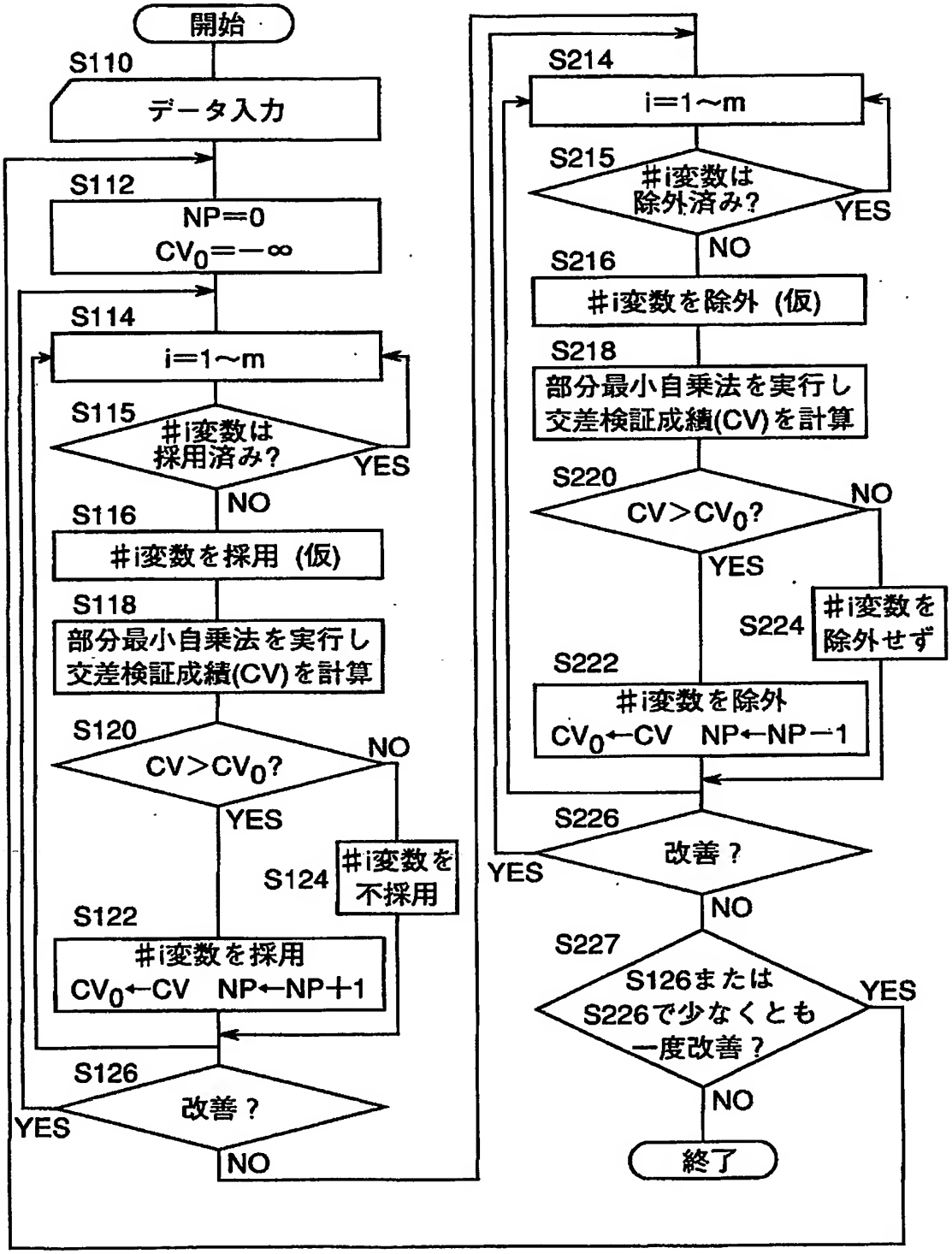
【図５】


開始
S210
データ入力
S212
NP=m
CV_0=−∞
S214
i=1～m
S215
#i変数は
採用済み?
YES
NO
S216
#i変数を除外 (仮)
S218
部分最小自乗法を実行し
交差検証成績(CV)を計算
S220
CV>CV_0?
NO
YES
S222
#i変数を除外
CV_0←CV NP←NP−1
S224
#i変数を
除外せず
S226
改善？
YES
NO
終了

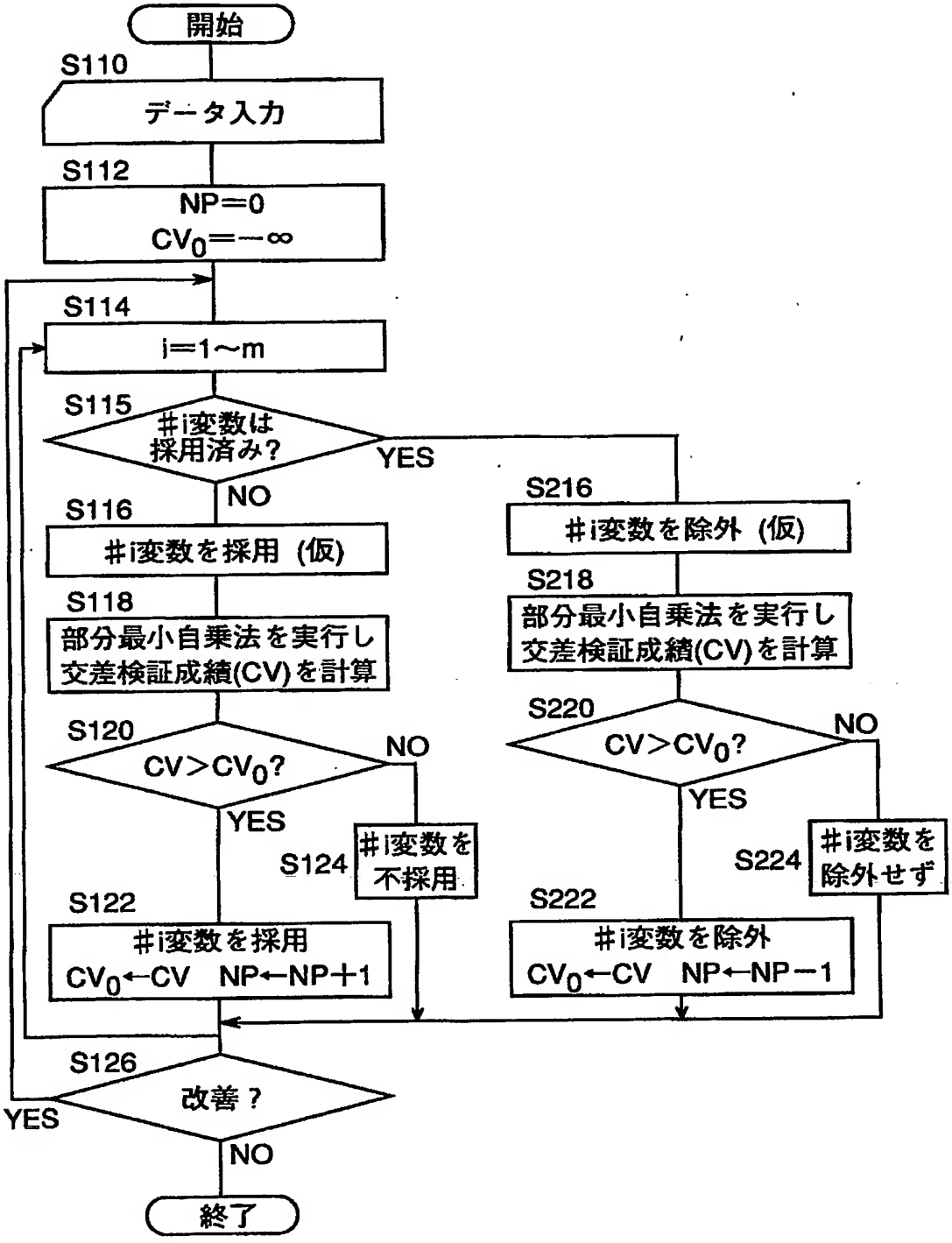
【図６】


開始
S110
データ入力
S112
NP=0
CV0=−∞
S114
i=1〜m
S115
#i変数は採用済み?
YES
NO
S116
#i変数を採用 (仮)
S118
部分最小自乗法を実行し交差検証成績(CV)を計算
S120
CV>CV0?
NO
YES
S124
#i変数を不採用
S122
#i変数を採用
CV0←CV NP←NP+1
S126
改善?
YES
NO
S214
i=1〜m
S215
#i変数は除外済み?
YES
NO
S216
#i変数を除外 (仮)
S218
部分最小自乗法を実行し交差検証成績(CV)を計算
S220
CV>CV0?
NO
YES
S224
#i変数を除外せず
S222
#i変数を除外
CV0←CV NP←NP+1
S226
改善?
YES
NO
終了

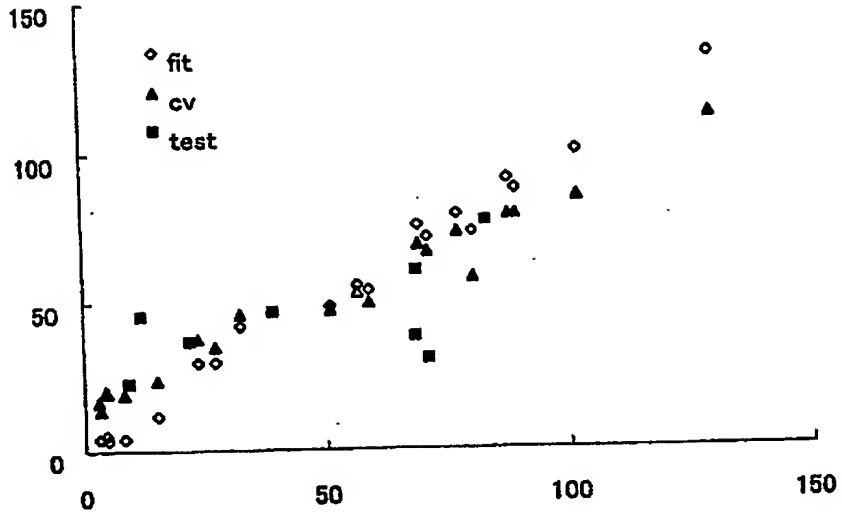
【図７】


開始
S110
データ入力
S112
NP=0
CV0=−∞
S114
i=1〜m
S115
#i変数は採用済み?
YES
NO
S116
#i変数を採用 (仮)
S118
部分最小自乗法を実行し交差検証成績(CV)を計算
S120
CV>CV0?
NO
YES
S124
#i変数を不採用
S122
#i変数を採用
CV0←CV NP←NP+1
S126
改善?
YES
NO
S214
i=1〜m
S215
#i変数は除外済み?
YES
NO
S216
#i変数を除外 (仮)
S218
部分最小自乗法を実行し交差検証成績(CV)を計算
S220
CV>CV0?
NO
YES
S224
#i変数を除外せず
S222
#i変数を除外
CV0←CV NP←NP−1
S226
改善?
YES
NO
S227
S126または S226で少なくとも一度改善?
YES
NO
終了

【図８】


開始
S110
データ入力
S112
NP=0
CV0=−∞
S114
i=1〜m
S115
#i変数は採用済み?
YES
NO
S116
#i変数を採用 (仮)
S118
部分最小自乗法を実行し交差検証成績(CV)を計算
S120
CV>CV0?
NO
YES
S124
#i変数を不採用
S122
#i変数を採用
CV0←CV NP←NP+1
S216
#i変数を除外 (仮)
S218
部分最小自乗法を実行し交差検証成績(CV)を計算
S220
CV>CV0?
NO
YES
S224
#i変数を除外せず
S222
#i変数を除外
CV0←CV NP←NP−1
S126
改善?
YES
NO
終了

【図９】



【書類名】　要約書

【要約】

【課題】　多変量の遺伝子発現情報の効果的な情報処理を提供する。

【解決手段】　生体の状態と複数の遺伝子発現の量および／または細胞内物質の量との相関モデルを決定するデータ解析において、

生体の状態を目的変数とし、複数（ｍ個）の遺伝子発現の量および／または細胞内物質の量を説明変数とするデータの集合において、データに含まれる説明変数を選択し、選択された説明変数と目的変数とを含む相関モデルについて交差検証成績を計算し、その結果を評価判定する。ここで、交差検証成績が改善しなくなるまで、説明変数の選択、交差検証成績の計算、その結果の評価判定を行い、部分最小自乗法モデルを決定する。

【選択図】図２

出証特２００３－３０３３８３５

出願人履歴情報

識別番号　〔０００００３５４〕

１．変更年月日　１９９３年　６月２１日
［変更理由］　住所変更
住　所　大阪府大阪市西区江戸堀一丁目３番１５号
氏　名　石原産業株式会社